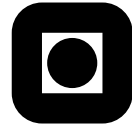


Oppgave 1



- a) `sx <- sd(x)`  
`sy <- sd(y)`  
`nx <- length(x)`  
`ny <- length(y)`  
`interval <- c(sx^2/(sy^2*qt(.975, df1=nx-1, df2=ny-1)), sx^2/(sy^2*qt(.025, df1=nx-1, df2=ny-1)))`  
`interval`
- b) `sqrt(interval)`

Oppgave 2

- a) Vektoren  $X_{AA}, X_{Aa}, X_{aa}$  er multinomisk fordelt med parametere  $n = 100$  og multinomiske sannsynligheter  $P_{AA}, P_{Aa}, P_{aa}$  lik frekvensen av genotypene i populasjonen utvalget er trukket fra.

```
dmultinom(c(24,54,24),size=100,prob=c(.25,.5,.25))
```

- b) Vi får  $\hat{p}_A = (2 \cdot 22 + 54)/(2 \cdot 100) = 0.49$ , som gir  $\hat{P}_{AA} = \hat{p}_A^2 = 0.49^2 = 0.2401$ ,  $\hat{P}_{Aa} = 2\hat{p}_A(1 - \hat{p}_A) = 0.4998$  and  $\hat{P}_{aa} = 0.2601$ .

- c) Vi har en situasjon hvor  $k = 3$  multinomiske sannsynligheter er funksjoner av  $s = 1$  parameter som her er estimert ved sannsynlighetsmaksimering.  $D$  er da tilnærmet kjikvadrat fordelt med  $k - 1 - s = 1$  frihetsgrad.

Dersom de observerte antallene av hver genotype  $X_{AA}, X_{Aa}, X_{aa}$  avviker mye fra forventet estimert antall under Hardy-Weinberg likevekt ( $H_0$ ) vil  $D$  ta en stor verdi. For signifikansnivå  $\alpha$  forkaster vi derfor  $H_0$  dersom  $D > \chi_\alpha^2 = 3.84$  (tabell). Gitt observert verdi  $D = 0.64$  beholder vi dermed  $H_0$ .

Testens p-verdi kan beregnes med uttrykket

```
pchisq(0.64, df=1, lower.tail=FALSE)
```

- d) Estimater av genotypfrekvensen  $\hat{P}_{aa}$  er funksjonen  $\hat{P}_{aa} = f(\hat{p}_A) = (1 - \hat{p}_A)^2$  av estimatet av allelfrekvensen  $A$ . Bruker vi delta-metoden finner vi at

$$\text{Var}(\hat{P}_{aa}) = \left( \frac{df}{dp_A} \right)^2 \text{Var}(\hat{p}_A) = (-2(1 - \hat{p}_A))^2 \text{Var}(\hat{p}_A) = 4(1 - \hat{p}_A)^2 \text{Var}(\hat{p}_A). \quad (1)$$

Et estimat av  $SE(\hat{P}_{aa})$  blir dermed  $\sqrt{4 \cdot 0.51^2 \cdot 0.035^2} = 0.0357$ .

- e) Lar vi  $i = 1, 2, 3$  betegne genotypene  $aa, aA$  og  $AA$  er modellen at

$$Y = \mu + \alpha_i + e$$

hvor residualledet  $e \sim N(0, \sigma^2)$  og uavhengig mellom ulike observasjoner. Parameterestimaterne er  $\hat{\mu} = 170.95$ ,  $\hat{\alpha}_2 = -22.97$ ,  $\hat{\alpha}_3 = -19.18$  og  $\hat{\sigma} = 11.97$  (effekten  $\alpha_1$  av å tilhøre kategori  $i = 1$  er satt lik 0 for å gjøre modellen identifiserbar).

Forventet hemoglobinnivå for de tre genotypene blir

|    |                              |
|----|------------------------------|
| aa | $170.95 + 0 = 170.95$        |
| Aa | $170.95 + (-22.77) = 148.18$ |
| AA | $170.95 + (-19.18) = 151.77$ |

Null- og alternativ hypotese i linjene merket (i) og (ii) er henholdsvis  $H_0 : \alpha_2 = 0$  vs  $H_1 : \alpha_2 \neq 0$  og  $H_0 : \alpha_3 = 0$  vs  $H_1 : \alpha_3 \neq 0$ , altså om det er noen forskjell mellom genotype *Aa* og kontrollgruppen *aa* (i) og mellom *AA* og *aa* (ii).

- f) Modellenen **recessive** (som er spiller rollen som  $H_0$  er nøstet i **fullmodel** (her  $H_1$ ) fordi denne framkommer som spesialtilfelle vi får ved å sette  $\alpha_2 = \alpha_3$ . F-testen av  $H_0$  vs  $H_1$  utført ved hjelp `anova(recessive, fullmodel, test=F)` indikerer at  $H_0$  (ingen forskjell mellom hemoglobinnivå for genotypene *Aa* og *AA*) ikke kan forkastes til fordel for  $H_1$  og vi beholder derfor modellen **recessive** som modell.

- g) Modellen er her at

$$Y = \mu + \beta \text{dosage} + e.$$

Igjen gjelder de samme antakelsene for residualledet  $e$  som i punkt e).

Denne modellen er også nøstet i **fullmodel** siden modellen framkommer som spesialtilfellet  $\alpha_2 = \beta \cdot 1$  og  $\alpha_3 = \beta \cdot 2$  men den er ikke noe spesialtilfelle av modellen **recessive** (siden sammenhengen mellom dosage og responsvariabelen ikke er lineær under dette modellalternativet).

F-testen indikerer at **linear** kan forkastes til fordel for **fullmodel** men siden denne ikke passer bedre til datene enn den enklere modellen **recessive** velger vi **recessive** som endelig valgt modell.

Dette stemmer bra med gjennomsnittsverdiene innen hver av de tre kategoriene i vedlagt figur.

- h) Under valgt modell finner vi ved bruk av tabell at

$$P(Y > 170|aa) = P\left(Z > \frac{170 - (170.75 + 0)}{11.28}\right) = P(Z > -0.0664) = 1 - P(Z \leq -0.0664) = 0.52$$

$$P(Y > 170|Aa) = P\left(Z > \frac{170 - (170.75 + (-19.19))}{11.28}\right) = \dots = 0.051$$

$$P(Y > 170|AA) = P(Y > 170|Aa) = 0.051$$

Anvender vi så Bayes teorem får vi

$$\begin{aligned} P(aa|Y > 170) &= \frac{P(Y > 170|aa)P(aa)}{P(Y > 170|aa)P(aa) + P(Y > 170|Aa)P(Aa) + P(Y > 170|AA)P(AA)} \\ &= \frac{0.52 \cdot 0.2601}{0.52 \cdot 0.2601 + 0.051 \cdot 0.4998 + 0.051 \cdot 0.2401} \\ &= 0.781 \end{aligned}$$